USING MACHINE LEARNING TO TRACK OBJECTS ACROSS CAMERAS

Yuewei Lin
Brookhaven National Laboratory
Upton, NY, USA

Xiaoqian Zhang
Brookhaven National Laboratory
Upton, NY, USA

Ji Hwan Park
Brookhaven National Laboratory
Upton, NY, USA

Shinjae Yoo
Brookhaven National Laboratory
Upton, NY, USA

Yonggang Cui
Brookhaven National Laboratory
Upton, NY, USA

Maikael Thomas
International Atomic Energy Agency
Vienna, Austria

Martin Moeslinger
International Atomic Energy Agency
Vienna, Austria

ABSTRACT

Video surveillance is one of the most important technologies used by the International Atomic Energy Agency in international safeguards. At large, complicated facilities, multiple surveillance cameras are deployed to monitor the transfer of safeguards-relevant objects across the site. During inspections, all surveillance videos are reviewed to ensure the objects are not manipulated or diverted during transfer, a laborious, time-consuming task. This work describes using deep machine learning algorithms to track objects automatically across multiple cameras, greatly improving the efficiency of the review process. The fundamental problem in this object tracking task across multiple cameras is how to associate the same object, which may show extreme intra-class variations, such as viewpoints, occlusions, and various scales, in different and even non-overlapped cameras. Object re-identification (Re-ID) in nuclear facility video surveillance is even more challenging than classic person or vehicle Re-ID problems because different instances in the same category may display an identical appearance. One observation from nuclear facility surveillance videos is that all objects must be carted (e.g., via forklift) to move. Therefore, the spatial context information of an object, which provides the feature from the carrier, is critical for the object Re-ID task. This work proposes a two-stream convolutional neural networks model that takes features of objects and their surrounding regions into account. Moreover, the custom videos usually are gleaned from different scenes from the training data, which may have extreme variations in illumination changes and/or cluttered backgrounds. Directly applying the trained model to custom videos will dramatically decrease the performance. To tackle this problem, an advanced domain adaptation technique is proposed to mitigate the gap between the data taken from different scenes. The proposed framework will track objects of interest across a nuclear complex. The resulting tracks can be used in further analyses, such as event/activity recognition, anomaly detection, etc.

BNL-221782-2021-COPA

INTRODUCTION

Video surveillance is used for a variety of purposes, including physical security monitoring for sensitive facilities, remote monitoring and verification of nonproliferation treaties, and nuclear safeguards. It is one of the primary technologies used by the International Atomic Energy Agency (IAEA) in the implementation of international safeguards. Currently, IAEA inspectors manually review all surveillance videos to confirm that objects containing nuclear materials are not manipulated or diverted during transfer or storage and to confirm the absence of undeclared activities; this manual review is a laborious, time-consuming task. A machine-learning-based object tracking method is being developed by Brookhaven National Laboratory (BNL) to help IAEA inspectors review the large number of surveillance videos from the nuclear facilities. As nuclear facility workplaces usually are quite large, multiple cameras are used to monitor each facility. Therefore, this work is divided into two sub-tasks. The first one is tracking objects within the field of view of a single camera – traditional object tracking task in machine learning, and the second involves tracking objects across cameras by associating the same object in different cameras – the so-called *object re-identification* (Re-ID) problem.

The Re-ID task has been studied extensively in two applications: person Re-ID [1] and vehicle Re-ID [2]. It is challenging because of the large intra-class variation, i.e., the same object may appear quite differently in various cameras in terms of the viewpoints, occlusions, and illuminations. In even worse cases, the same object seen from different viewpoints or under different illumination conditions may appear to differ even more than different objects do under the same conditions. In this project, the object Re-ID task is more challenging than the person or vehicle Re-IDs as different instances of the same object may appear identical. Figure 1 shows two example images. The boxes outlined in red contain the same object, while the boxes in blue have different objects. However, without color box labeling, it is difficult to discern the differences visually.

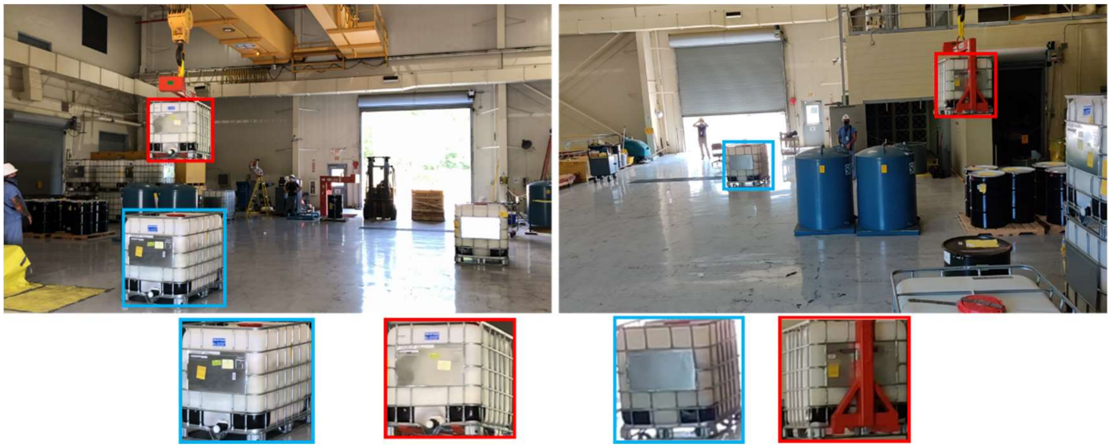

Figure 1. Examples of the images and objects that detail the complexity in object Re-ID.

Fortunately, the object Re-ID task has a unique feature compared to the person and vehicle Re-ID tasks. Both people and vehicles can move “by themselves” without any carriers, but the objects in nuclear facilities usually are bulky and heavy and must be moved by either a crane, forklift, or truck. With this assumption, the spatial context information, which provides the carrier information, becomes critical to the object Re-ID task. Therefore, a two-stream machine learning model is proposed here that takes pairs of image patches – object and surrounding – and determines whether they belong to the same object by considering features of both the object and its surroundings.

DATA COLLECTION

To train and demonstrate the proposed model's effectiveness, both indoor and outdoor test datasets were collected at Brookhaven National Laboratory's Waste Management Facility. Figure 2 shows the indoor data collection setting with five different cameras to cover the scene. There were five objects in total. The object carried by a crane (referred to as Box A) and the object carried by a forklift (referred to as Box B), moving along the two dotted lines, were considered as objects of interest. Three others, referred to as Box 1, 2, and 3, were standing on the floor, and treated as objects of non-interest (NI Box). Cameras could record both Boxes A and B at certain times. For the outdoor scenario, a box, referred to as Box C, was carried by a forklift moving into and out of the building. Figure 3 shows the sample images of Boxes A, B, and C. In the experiments, Boxes B and C were treated as the same object but in different scenarios.

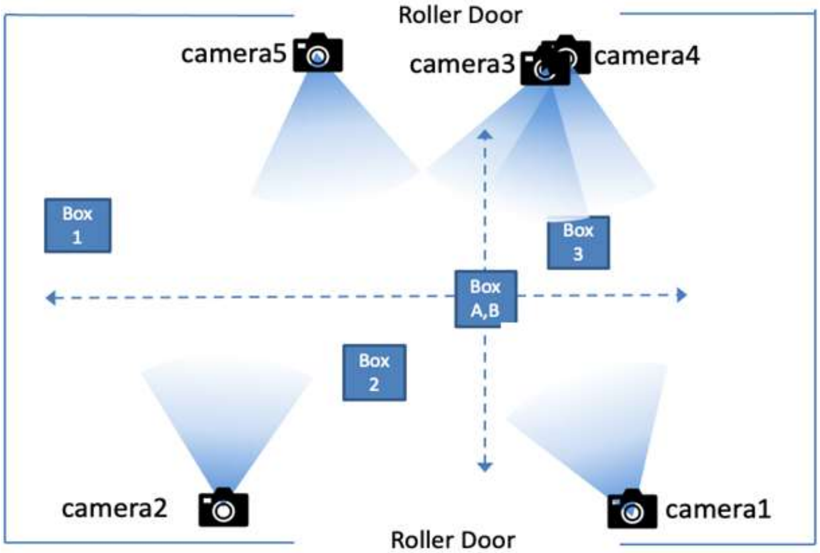

Figure 2. Data collection environment.

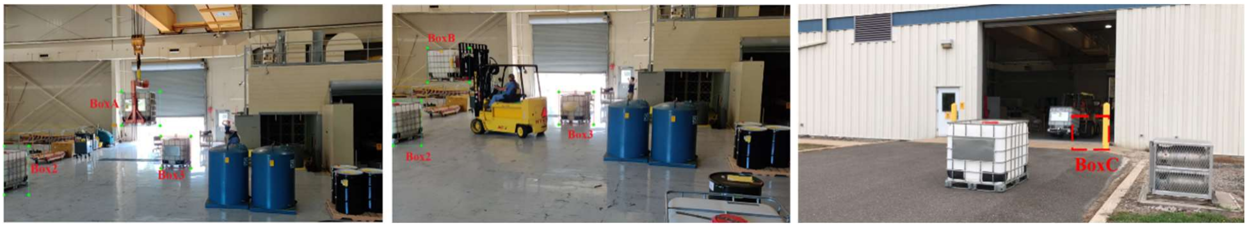

Figure 3. Sample images for different boxes in specific scenarios.

DEEP LEARNING MODEL

Object Re-ID with a Two-stream Deep Convolutional Neural Network Model

As mentioned, a two-stream deep convolutional neural network (CNN) model, motivated by [3], is used to tackle the object Re-ID problem. The object Re-ID problem is considered as the binary classification task, i.e., with input consisting of a pair of object images, the model predicts either "matching" (the same object) or "non-matching" (different objects). As shown in Figure 4, the model has two streams, one for the object images and the other for their surrounding regions. The object stream takes two object images and extracts their features by using the same CNN model, while the surrounding stream also takes two surrounding images and extracts their features by using another CNN model. The features from the object and surrounding streams are concatenated together then fed into a binary classifier.

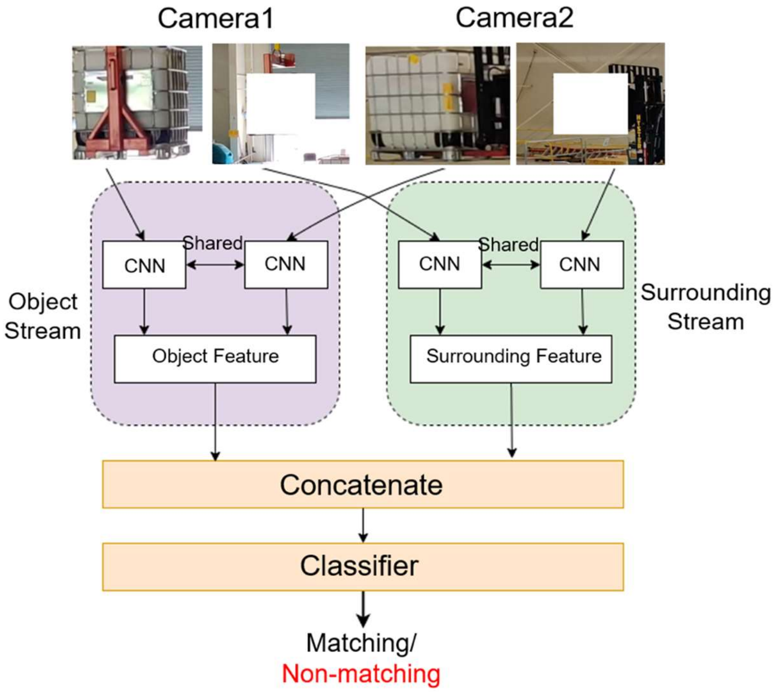

Figure 4. Illustration of the two-stream deep CNN model.

Object Re-ID with Domain Adaptation

In this project, the images taken from different cameras may have distinct illumination and blur conditions and background clutter as seen in the datasets discussed previously. Directly applying the model trained from one scenario (e.g., indoor), referred to as the *source domain*, to another (e.g., outdoor), or the *target domain*, usually results in significantly degraded performance (described in Section 4.3), while labeling data in the target domain as training samples would be time-consuming. To solve the scenario difference problem without labeling any data in the target domain, the domain adaptation technique can be used by aligning the source and target distributions in feature space. There are two approaches commonly used to project data in both source and target domains to a common feature space, where source and target feature distributions are similar.

The first approach is the Maximum Mean Discrepancy (MMD) [4]. MMD is used to measure the difference between two distributions. Thus, minimizing the MMD of the distributions of the source and target data could reduce the distributional difference between two domains.

MMD is defined as Eq. (1), which is the distance of the mean of two distributions,

$$\mathrm{MMD} = ||\frac{1}{n_s}\sum_{\boldsymbol{x}_i \in \boldsymbol{X}_s} \boldsymbol{A}^\top \boldsymbol{x}_i - \frac{1}{n_t}\sum_{\boldsymbol{x}_j \in \boldsymbol{X}_t} \boldsymbol{A}^\top \boldsymbol{x}_j||_2^2 ,$$

where X_s and X_t denote the source and target domain and n_s and n_t represent the data numbers in the source and target domain, respectively. A^{T} indicates a mapping from the original image space to the feature space, which is a feature extraction process. In this study, it is a CNN model.

The second approach is adversarial learning [5]. Much like methods used in generative adversarial networks (GAN), a discriminator is introduced to determine which domain the feature comes from, while the feature extractor tries to generate features to fool the discriminator, i.e., the feature extractor attempts to extract the features that are domain invariant. Once trained, the features from different domains are well aligned.

While conventional MMD works well on domain adaptation for the standard image classification problem, it has difficulties handling the open set object Re-ID problem [6]. Therefore, we convert the original feature space to be a so-called *dissimilarity space*. Specifically, it is generated by calculating the within-class and between-class distances. In this study, by tracking the objects within a video, we can easily obtain the same instance to calculate the within-class distances. All other instance pairs are used to calculate the between-class distances.

For the labeled source domain, we employ standard supervised learning with cross entropy loss function and standard metric learning with triplet loss. For the unlabeled target domain, we use MMD minimization in both original feature space and dissimilarity space. The MMD minimization is used to align the distributions in different domains.

EXPERIMENTAL RESULTS

Experimental Setting

As described in Section 2, there are two objects of interest in the datasets. The first one (Box A) was carried by an indoor crane, and the second one was carried by a forklift in both the indoor scene (Box B) and the outdoor scene (Box C). All other boxes are non-interest objects. In this study, it is assumed that all the objects already have been detected, e.g., by the You-Only-Look-Once (YOLO) object detection algorithm introduced to safeguards applications in the early work of this project [7]. Thus, all objects and their corresponding surrounding regions are labeled and ready to use.

We evaluate two different settings: the inner-object and across-object settings. For the inner object setting, all training data is from one object of interest, e.g., Box A, while all testing data also are from Box A. In contrast, for the across-object setting, the model is trained and tested with different objects of interest. For example, if Box A is used for training, then the positive pair is two images of Box A, while the negative pair is one image of Box A and one image of another object. The testing pairs are all Box B. During training, the model has never seen Box B with the forklift, but the trained model will be tested on Box B. Therefore, the across-object setting shows the model's generalization ability. Along with the inner- and across-object settings, we also evaluate the inner- and across-domains settings, i.e., the objects may be from a different scenario.

Evaluation Metrics

As in other classification tasks, recall, precision, and F1-score are used as evaluation metrics in this study. As shown in Figure 5, given a pair of the same objects, if the model predicts as positive, it is considered as a true positive. Otherwise, it is treated as a false negative. Similarly, given a pair of different objects, if the model predicts as positive, it is considered as a false positive; otherwise, it is treated as a true negative. Recall and precision are defined as

$$Recall = \frac{True\ Positive}{True\ Positive{+}Fa \quad Negative} \quad Precision = \frac{True\ Positive}{True\ Positive{+}Fal \quad Prositive} \ .$$

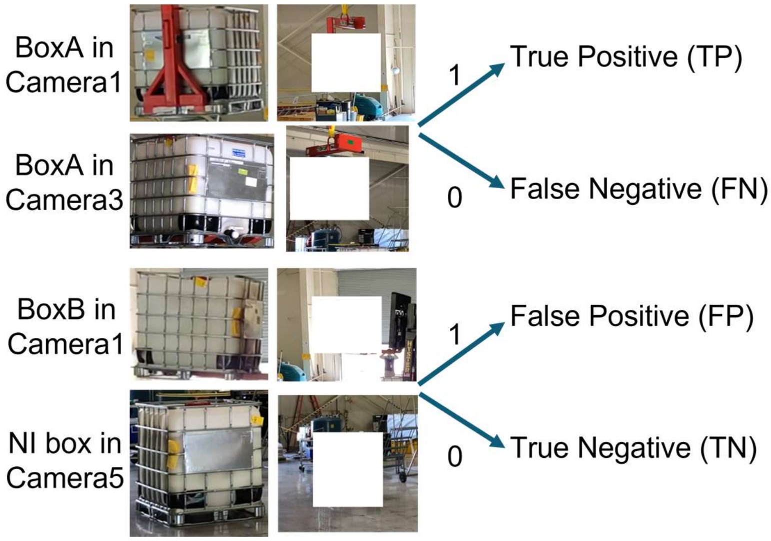

Figure 5. Illustrations of true positive, false negative, false positive and true negative.

Generally, high recall means given a pair of the same objects, there is a small chance it will be classified as negative. High precision means given a pair of different objects, there is a small chance it will be classified as positive. The F1 score is the harmonic mean of recall and precision, which is a balance of them, defined as follows:

$$F1\ score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \ .$$

Results

For the indoor setting, the first and third columns in Table 1 show the inner object setting results, which are almost perfect. The second and fourth columns show the across-object settings, where the precisions are reasonable, but recall is lower than those for the inner-object settings.

Table 1. Results of different settings.

	A->A	A->B	B->B	B->A	A->C	B->C	C->A	C->B
Precision	100	96.66	100	99.95	67.42	50.30	58.98	57.26
Recall	100	84.67	99.88	76.81	12.00	24.80	52.20	63.50
F-1	100	90.27	99.95	86.87	20.37	33.22	55.38	60.22

For the across scenarios/domains settings, the fifth and seventh columns of Table 1 show the results of the across-object and across-domain settings, while the sixth and eighth columns depict the results of the inner-object and across-domain settings. Its accuracy dropped considerably compared to the inner-domain settings.

To improve performance of the across-domain settings, a small number of target images were added into the training set. For example, if the model is trained on Box A (source data) and a small number of Box C (target data), then testing is done on Box C and the experiment is noted as Box A -> Box C. The left sub-figure in Figure 6 shows the (Box A->Box C) performance with a different number of target data added, while the right sub-figure depicts the (Box B->Box C) performance. With more target data added, the precision usually remains the same (or slightly drops), while recall increases. With the help of additional target data, the F1-score increases from 20.37% (Box A->Box C) and 55.38% (Box A->Box C) to 66% and 70%, respectively.

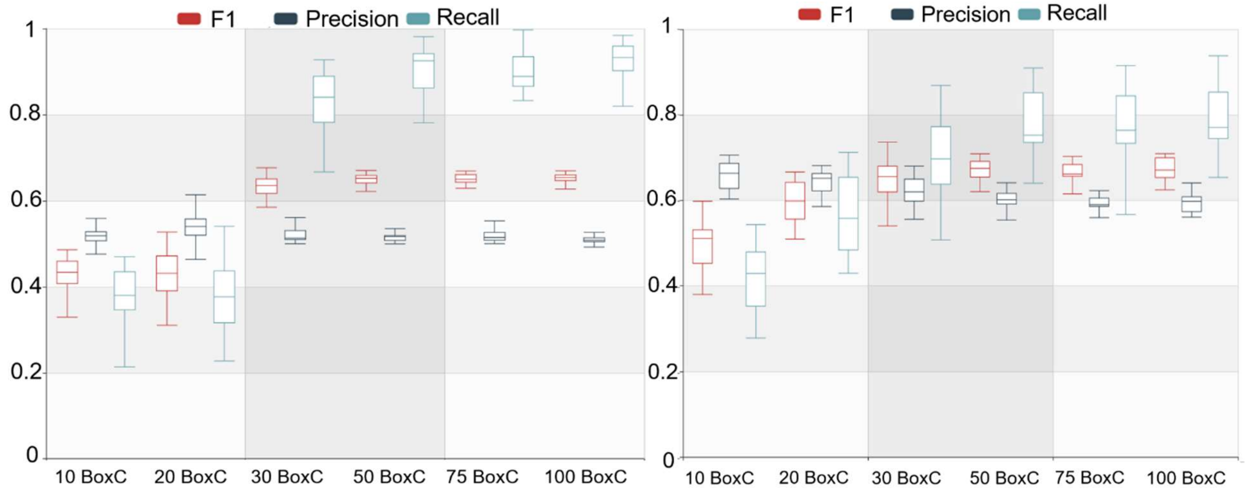

Figure 6. The Box A->Box C (left) and Box B->Box C (right) performance of varying amounts of additional target data appended to the training set without data augmentation techniques.

To further improve the performance, data augmentation was implemented in the model to mimic the scenario difference. Specifically, two basic augmentation techniques, rotation, and illumination adjustment, were added to the model. Figure 7 illustrates the results of the Box B->Box C setting with data augmentation. It shows an F1 score of 74% compared to 70% without data augmentation. More importantly, the recall value improved significantly, especially when fewer images of Box C were used in training.

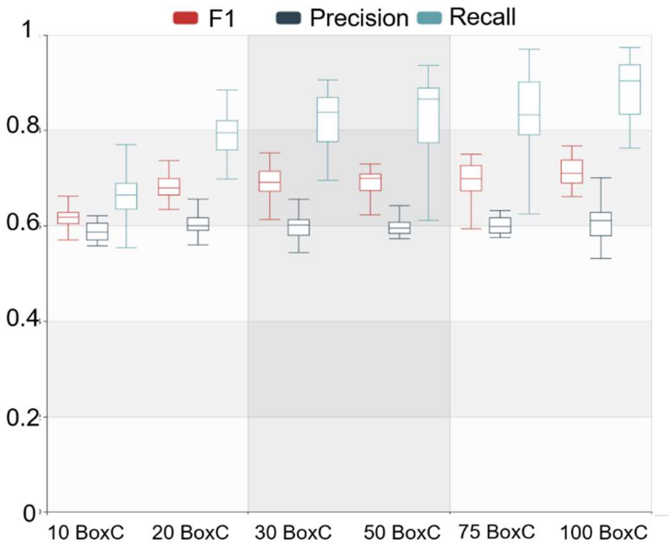

Figure 7. The Box B->Box C performance of different additional target data appended to the training set with data augmentation techniques.

Object Re-ID with the domain adaptation technique is being evaluated. Here, the dataset is split into indoor and outdoor scenes, and the model is trained on the outdoor scene with 982 images and tested on 283 indoor images. The evaluation metric used here is the rank accuracy, which is widely used in image retrieval tasks. Given an image, the model will first rank all other images based on their distances to the selected one. Rank-k accuracy is calculated based on top *k* closest images. The rank-1 and rank-5 accuracies are 83.3% and 100%, respectively. The results are preliminary, and we will utilize the data augmentation and hyper-parameter optimization techniques to further improve the accuracy in the future.

CONCLUSIONS

Deep machine learning methods were proposed to automatically track objects across multiple cameras, especially in relation to safeguarding sensitive facilities. Specifically, a deep CNN model was developed to extract the context information by taking pairs of image patches, from both images of objects and their surrounding regions. The test results with different experimental settings, including inner/across objects and inner/across domains, have shown that this deep CNN model is effective in re-identifying objects across multiple cameras. However, the method requires significant effort in labeling the training images, which is a common issue in supervised ML methods. To solve the issue, domain adaptation techniques were proposed to improve the performance of the across-domain settings without any labeled target data available. Test results on the domain adaptation methods will be reported in the future.

ACKNOWLEDGEMENTS

The authors are grateful to Glen Todzia, Edward Richards and John Aloi Jr. at BNL for their efforts in arranging data collection.

The work presented in this paper was funded by the National Nuclear Security Administration of the Department of Energy, Office of International Nuclear Safeguards (NA-241). The manuscript has been authored by Brookhaven National Laboratory managed by Brookhaven Science Associates, LLC under Contract No. DE-SC0012704 with the U.S. Department of Energy.

REFERENCES

1. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L. and Hoi, S.C. 2020. Deep learning for person re-identification: A survey and outlook, arXiv preprint arXiv:2001.04193.
2. He, B., Li, J., Zhao, Y. and Tian, Y. 2019. Part-regularized near-duplicate vehicle re-identification, IEEE Conference on Computer Vision and Pattern Recognition.
3. Oliveira, I. O., Fonseca, K. V. and Minetto, R. 2019. A two-stream Siamese neural network for vehicle re-identification by using non-overlapping cameras, IEEE International Conference on Image Processing.
4. Wang, W., Li, H., Ding, Z. and Wang, Z. 2020. Rethink Maximum Mean Discrepancy for Domain Adaptation (arXiv preprint), arXiv:2007.00689.
5. Tang H. and Jia, K. 2020. Discriminative Adversarial Domain Adaptation, AAAI Proceeding.
6. Mekhazni, D., Bhuiyan, A., Ekladious, G. and Granger, E. 2020. Unsupervised domain adaptation in the dissimilarity space for person re-identification, European Conference on Computer Vision.
7. Cui, Y., Gastelum, Z., Ren, Y., Yoo, S., Lin, Y., Smith, M. R., Thomas, M. A., and Stern, W. 2018. Using deep machine learning to conduct object-based identification and motion detection on safeguards video surveillance, Proceedings of 2018 IAEA Symposium on International Safeguards: Building Future Safeguards Capabilities, Vienna, Austria.